



(19)

(11) Publication number:

08030558 A

Generated Document.

PATENT ABSTRACTS OF JAPAN

(21) Application number: 06167902

(51) Intl. Cl.: G06F 15/16

(22) Application date: 20.07.94

(30) Priority:

(43) Date of application publication: 02.02.96

(84) Designated contracting states:

(71) Applicant: FUJITSU LTD

(72) Inventor: UKAI TAKANORI
UEDA HARUYASU

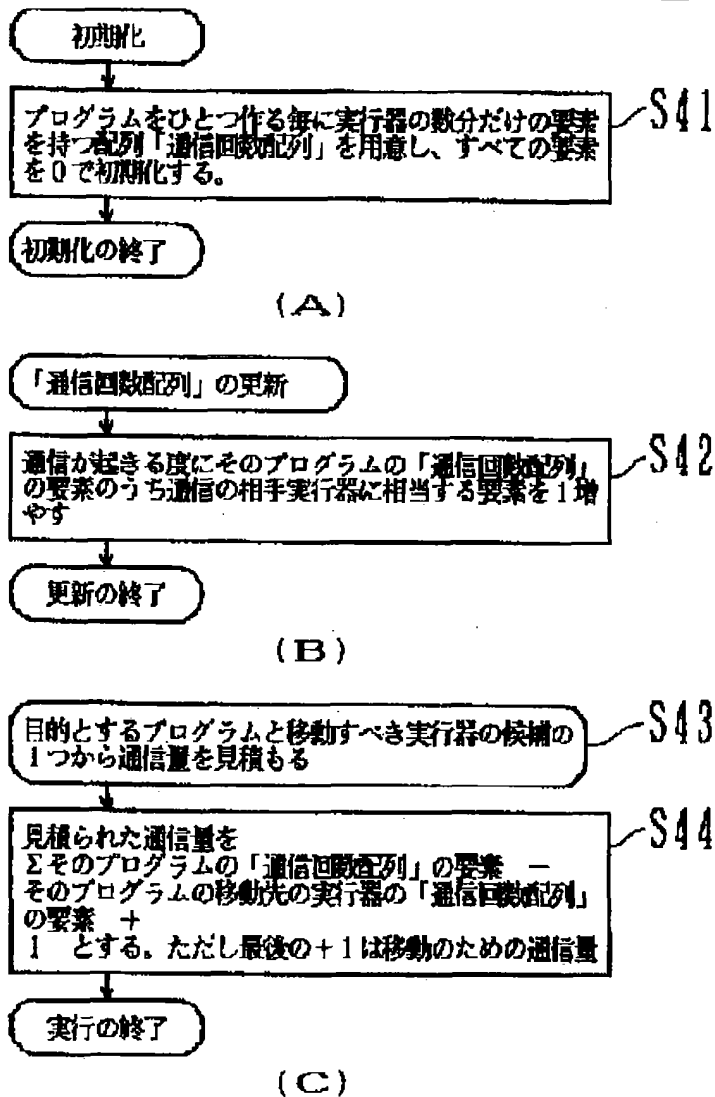
(74) Representative:

(54) LOAD DISTRIBUTING METHOD IN COMPUTER SYSTEM AND COMPUTER SYSTEM UTILIZING THE METHOD

(57) Abstract:

PURPOSE: To improve the processing efficiency of the whole computer system by properly distributing load.

CONSTITUTION: An execution device (i) adds '1' to the value of a corresponding program element in a communication frequency array in each communication of a program in execution with each execution device (S42). The volume of communication between a program (m) to be a moving candidate and an execution device (n) to be a candidate to which the program (m) is to be moved is estimated (S43). Communication frequency with the execution device (n) is subtracted from the sum of values of respective arrays in the program (m) to be the moving candidate of the communication frequency array, i.e., the sum of communication frequency, '1' is added to the subtracted result and the added result is set up as the volume of communication after moving the program (m). On the other hand, the sum of elements of an execution device other than the execution device (i) is set up as the volume of communication obtained before the movement of the program (m). Then the communication volume before movement and the communication volume after movement are respectively multiplied by a communication time coefficient '1' to find out communication time before movement and communication time after movement, and when a combination between a program and an execution device by which the communication time after movement is shorter than the communication time before movement exists, the program is allocated to the execution device.



COPYRIGHT: (C)1996,JPO

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平8-30558

(43) 公開日 平成8年(1996)2月2日

(51) Int.Cl.⁶

G 0 6 F 15/16

識別記号

3 7 0 N

片内整理番号

F I

技術表示箇所

審査請求 未請求 請求項の数 6 O L (全 8 頁)

(21) 出願番号 特願平6-167902

(22) 出願日 平成6年(1994)7月20日

(71) 出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中1015番地

(72) 発明者 鶴飼 孝典

神奈川県川崎市中原区上小田中1015番地

富士通株式会社内

(72) 発明者 上田 晴康

神奈川県川崎市中原区上小田中1015番地

富士通株式会社内

(74) 代理人 弁理士 大菅 義之 (外1名)

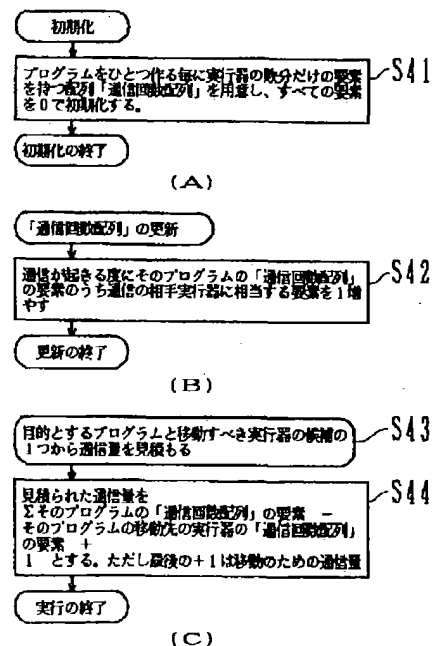
(54) 【発明の名称】 計算機システムにおける負荷分散方法及びそれを利用した計算機システム

(57) 【要約】

【目的】 本発明の目的は、負荷を適正に分配することにより計算機システム全体の処理効率を向上させることである。

【構成】 実行器 i は、実行中のプログラムが各実行器と通信を行う度に、通信回数配列の該当するプログラムの要素の値を「1」インクリメントする(図6(B)のS42)。移動候補のプログラム m とそのプログラム m の移動先候補の実行器 n との間の通信量を見積もる(同図(C)のS43)。そして、通信回数配列の移動候補のプログラム m の各配列の値の総和、即ち通信回数の総和から、移動先候補の実行器 n との通信回数を減算し、減算結果に「1」を加えたものを、プログラム m の移動後通信量とする。また、実行器 i 以外の他の実行器の要素の総和をプログラム m の移動前通信量とする。そして、移動前通信量及び移動後通信量に通信時間係数として「1」を乗算して移動前通信時間及び移動通信時間を求め、移動後通信時間が移動前通信時間より短くなるプログラムと実行器の組み合わせが存在したなら、その実行器にプログラムを割り当てる。

通信量見積り処理のフローチャート



1

【特許請求の範囲】

【請求項1】 複数のプログラムをそれぞれ実行する複数の実行器と、該複数の実行器を制御する制御器とからなる計算機システムにおいて、

前記実行器が、

実行すべきプログラムと自己の実行器を含む各実行器との間の通信量を見積もる通信量見積り手段と、

前記通信量見積り手段によって見積もられた通信量に基づいて、前記プログラムを自己の実行器で実行した場合

の該プログラムと他の実行器との間の移動前通信時間と、該プログラムを他の実行器に移動させた場合の該プログラムと移動先実行器以外の他の実行器との間の移動後通信時間とを算出する通信時間算出手段と、

前記通信時間算出手段によって算出された移動後通信時間が移動前通信時間より短くなるプログラムと実行器との組み合わせを抽出し、該抽出した実行器に該プログラムを割り当てるように前記制御器に通知する負荷割り当て手段とを備えることを特徴とする計算機システム。

【請求項2】 前記通信量見積り手段は、前記プログラムと自己の実行器を含む各実行器との間の過去の通信量を記憶する通信量記憶手段を有し、

前記通信時間算出手段は、前記記憶手段に記憶されている過去の通信量から前記移動前通信時間と前記移動後通信時間を算出することを特徴とする請求項1記載の計算機システム。

【請求項3】 前記通信量見積り手段は、前記プログラムと前記各実行器との通信回数を記憶する記憶手段を有し、

前記通信時間算出手段は、前記記憶手段に記憶されている通信回数と、予め定められている各実行器との間の通信時間係数とから前記プログラムと前記各実行器との間の通信時間を算出することを特徴とする請求項1記載の計算機システム。

【請求項4】 前記通信時間算出手段は、全ての実行器または前記制御器から指示された処理能力に余裕のある実行器との間で通信を行い、そのときの通信時間から前記プログラムと前記各実行器との間の通信時間係数を定めることを特徴とする請求項1記載の計算機システム。

【請求項5】 前記制御器は、処理能力に余裕のある実行器を、プログラムの移動を希望する実行器に通知する通知手段を有することを特徴とする請求項1記載の計算機システム。

【請求項6】 複数のプログラムをそれぞれ実行する複数の実行器と、該複数の実行器を制御する制御器とからなる計算機システムの負荷分散方法において、

実行すべきプログラムと各実行器との間の通信量を記憶し、

該記憶した通信量に基づいて、前記プログラムを自己の実行器で実行した場合の該プログラムと他の実行器との間の移動前通信時間と、該プログラムを処理能力に余裕

2

のある他の実行器に移動させた場合の該プログラムと移動先実行器以外の実行器との間の移動後通信時間とを算出し、

前記移動後通信時間が移動前通信時間より短くなるプログラムと実行器との組み合わせを抽出し、該抽出した実行器に該プログラムを割り当てることを特徴とする計算機システムの負荷分散方法。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、ネットワークで接続された複数の計算機の負荷を適正に分散する負荷分散方法及びその負荷分散方法を適用した計算機システムに関する。

【0002】

【従来技術】複数の計算機がネットワークで接続されたシステムでは、システム全体の処理時間を短くするため各計算機の負荷分担を適正化することが要求される。

【0003】負荷の適正化の方法として、各計算機の負荷が平均化されるようにする方法、あるいは負荷の最も軽い計算機に新たな仕事を分担させる方法等が考えられている。

【0004】

【発明が解決しようとする課題】ところで、上述した負荷分散方法では、プログラムを負荷の軽い計算機に割り当てるためにはプログラムをその計算機に転送する必要があり、計算機間のデータの転送時間がプログラムの実行時間に比較して長い場合には、単純に負荷を平均化しただけでは計算機システム全体の処理効率を向上させることはできない。特に並列計算機システムでは、計算機の計算時間に比べて計算機間の通信時間が長いので、単に負荷を均等に割り当てたのでは計算機システム全体の効率を向上させることができないという問題点があった。

【0005】本発明の課題は、計算機システム全体の処理効率を向上させることである。

【0006】

【課題を解決するための手段】図1は、本発明の負荷分散方法を適用した計算機システムの原理ブロック図である。複数のプログラムをそれぞれ実行する複数の実行器と、該複数の実行器を制御する制御器とからなる計算機システムにおいて、実行器は、通信量見積り手段と、通信時間算出手段と、負荷割り当て手段とを有している。

【0007】通信量見積り手段は、実行すべきプログラムと自己の実行器を含む各実行器との間の通信量を見積もる。この通信量見積り手段は、例えば実行すべき複数のプログラムと自己の実行器を含む各実行器との過去の通信量（通信回数、データ量等）を記憶し、その記憶してある通信量をプログラムと各実行器との間の通信量とする。

【0008】通信時間算出手段は、通信量見積り手段に

よって見積もられた通信量に基づいて、プログラムを自己の実行器で実行した場合のそのプログラムと他の実行器との間の移動前通信時間と、そのプログラムを他の実行器に移動させた場合のプログラムと移動先実行器以外の実行器との間の移動後通信時間とを算出する。

【0009】この通信時間算出手段は、例えばプログラムの実行に先立って他の実行器との間で通信を行い、そのときの通信時間を各実行器の通信時間係数として求め、通信量見積り手段により見積もられた通信量にそれぞれの実行器の通信時間係数を乗算して、プログラムと各実行器との間の通信時間を算出する。

【0010】負荷割り当て手段は、通信時間算出手段によって算出された移動後通信時間が移動前通信時間より短くなるプログラムと実行器との組み合わせを抽出し、その抽出した実行器にプログラムを割り当てるように制御器に通知する。

【0011】

【作用】本発明の計算機システムでは、例えば、それぞれのプログラムと各実行器との間の通信量を記憶しておいて、あるプログラムを自己の実行器で実行した場合の各実行器との間の移動前通信時間と、そのプログラムを他の実行器に移動させた場合の各実行器との間の移動後通信時間とを求め、移動後通信時間が移動前通信時間より短くなるようにプログラムを他の実行器に移動させるようにしたので、実行器間の通信時間が実行器の計算時間より長い場合にも、負荷を適正に分散して、計算機システム全体の処理時間を短縮することができる。

【0012】

【実施例】以下、本発明の実施例を図面を参照しながら説明する。図2は、本発明の実施例の計算機システムの構成図である。この実施例の計算機システムは、ネットワークで接続されたUNIX環境上に構築されている。UNIX環境では、複数のプロセスが1つの計算機上で動作するので、計算機上で動作するこれらのプロセスを実行器とし、複数の実行器を制御する制御用プロセスを制御器としてもよいし、プロセッサエレメント(PE)とメモリ等からなるハードウェアで実行器及び制御器を構成してもよい。

【0013】計算機1は、実行器11と制御器14とを有し、計算機2は実行器12を、計算機3は実行器13をそれぞれ有している。特定の実行器11から他の実行器にプログラムを移動させる場合には、実行器11が移動させるプログラムを制御器14に送信し、制御器14がそのプログラムを処理能力に余裕のある実行器に送信するようになっている。なお、計算機1〜3はバスを介して接続されている。

【0014】各実行器11〜13は、複数のプログラムを時分割で実行することができるとともに、制御器14から処理能力に余裕のある実行器のリストを受け取る。各実行器11〜13は、実行しようとするプログラムの

数が自己の実行可能なプログラム数より多いときには、そのことを制御器14に通知する。そして、制御器14から処理能力に余裕のある実行器のリストが通知されたなら、実行器11〜13は、停止させたプログラムあるいはこれから実行しようとするプログラムを再開あるいは開始させるための情報を制御器14に送る。

【0015】また、実行器11〜13は、制御器14からプログラムを再開あるいは開始するための情報を受け取り、そのプログラムを再開する。制御器14は、プログラムの転送要求を実行器から受け取ると、処理能力に余裕のある実行器のリストを転送要求のあった実行器に送る。また、制御器14は、実行器から送られてくるプログラムを再開するための情報を受け取り、それらの情報をプログラムの移動先の実行器に送る。

【0016】次に、図3は、実行器11の構成図であり、他の実行器12、13も図3と同様な構成を有する。同図に示すように実行器11は、プロセッサエレメント(PE)21と、後述する通信時間見積りプログラム、通信量見積りプログラム等を記憶するメモリ22と、実行器11が処理可能なプログラム数の上限値を記憶する上限値メモリ23と、実行器11が実行する各プログラムと自己及び他の実行器との間の通信回数を記憶する通信回数テーブル24と、プログラムの実行順序が記憶されるキュー25とを有する。

【0017】例えば、通信回数テーブル24には、プログラムaと実行器11とのそれまでの通信回数として「3」、実行器12との通信回数として「2」、実行器13との通信回数として「1」が記憶されている。他のプログラムについても同様に各実行器との間の現在までの通信回数が記憶されている。

【0018】次に、実行器11と制御器14の動作を図4及び図5のフローチャートを参照して説明する。以下、実行器11を例にとり説明する。実行器11は、図4のステップS1で実行すべきプログラムをキュー25に登録する。そして、キュー25からプログラムを取り出し順次実行する。そして、時間がきたなら、ステップS3でプログラムを停止させ、プログラムカウンタとデータとプログラム本体を保存する。そして、次のステップS4で現在実行しているプログラムとこれから実行しようとするプログラムの数を調べる。このとき、実行すべきプログラムの数が自分が実行できるプログラム数を超えているときには、ステップS5で実行すべきプログラム数が自己の処理可能なプログラム数を超えていることを制御器14に通知する。

【0019】そして、制御器14からその時点で処理能力に余裕がある実行器のリストが送られてきたなら、ステップS6で現在実行中のプログラムを停止させ、そのプログラムの移動先の実行器を指定して、プログラムカウンタの値と、データと、プログラム本体とを制御器14に送り、キュー25からそのプログラムを削除する。

このとき、実行器11は、通信回数テーブル24に記憶されている通信回数から算出される各プログラムと各実行器との間の通信時間と、各プログラムを他の実行器に移動させた場合の移動後通信時間とを算出し、移動後通信時間が移動前通信時間より短くなるプログラムと実行器との組み合わせがあったなら、その実行器をプログラムの移動先として制御器14に指示する。

【0020】他方、制御器14から再開すべきプログラムのプログラムカウンタの値と、データと、プログラム本体とが送られてきた場合には、ステップS7でそれらのデータ及びプログラムを内部のメモリ22に格納すると共に、そのプログラムをキュー25に登録する。

【0021】制御器14は、図5のステップS21で実行器から送られてくる命令を待ち、例えば実行器iから命令を受け取ったなら、ステップS22でその命令が実行器iのプログラム数を知らせるものか否かを判別する。

【0022】ステップS22の判別で、受け取った命令が実行器iのプログラム数を知らせる命令であると判別されたときには、各実行器の状態を調べるために、先ずステップS23で実行器11が実行中のプログラム数を調べ、実行器11が空いているか否か、すなわち実行器11で新たなプログラムを実行可能か否かを判別する。この判別で実行器11に空きがあると判別されたときには、ステップS24に進み、命令を発行した実行器iに実行器11の処理能力に余裕があることを通知する。

【0023】ステップS23の判別で実行器11が空いていないと判別されたとき、またはステップS24の次には、ステップS25に進み実行器12が空いているか否か、すなわち実行器12で新たなプログラムを実行可能か否かを判別する。この判別で実行器12に空きがあると判別されたときには、ステップS26に進み、命令を発行した実行器iに実行器12の処理能力に余裕があることを通知する。

【0024】ステップS25の判別で実行器12が空いていないと判別されたとき、またはステップS26の次には、ステップS27に進み実行器13が空いているか否か、すなわち実行器13で新たなプログラムが実行可能か否かを判別する。この判別で実行器13に空きがあると判別されたときには、ステップS28に進み、命令を発行した実行器iに実行器13の処理能力に余裕があることを通知する。

【0025】上記の処理により、実行器i側では、どの実行器が空いているかを知ることができるので、後述する通信時間見積り処理等により実行中のプログラムを空いている実行器に移動させた場合の通信時間を算出し、通信時間の短くなる実行器にプログラムを移動させることができる。

【0026】ステップS22の判別で実行器iから送られてきた命令がプログラム数を知らせる命令でなかったときには、ステップS29に進み、プログラムを再開す

べき実行器を指定する命令が実行器iから送られてきたか否かを判別する。この判別で、プログラムを再開すべき実行器を指定する命令であると判別されたときには、次のステップS30で、実行器iから送られてきたプログラム情報を指定された実行器に送信する。

【0027】次に、実行器iで実行中のプログラムの通信量を見積もる通信量見積り処理（通信量見積り手段に対応する）を図6（A）、（B）、（C）のフローチャートを参照して説明する。

10 【0028】先ず、図6（A）のステップS41で、実行器iが実行中のプログラムと自己の実行器を含む各実行器との間の通信回数が記憶される通信回数配列を通信回数テーブル24に設け、その通信回数配列の全ての要素を「0」で初期化する。次に、同図（B）のステップS42で、実行器iで実行中のプログラムが各実行器と通信を行う度に、通信回数配列の該当するプログラムの各要素の中で通信の相手先の実行器に対応する要素の値を「1」インクリメントする。上記の処理により、例えば、プログラムaが実行器11と通信した場合には、図3の通信回数テーブル24のプログラムaと実行器11との要素の値が「3」から「4」に変更される。

【0029】次に同図（C）のステップS43で、移動候補のプログラム（これを、プログラムmとする）とそのプログラムmの移動先候補の実行器（これを、実行器nとする）と間の通信量を見積もる。具体的には、通信回数配列の移動候補のプログラムmと移動先実行器nとの要素の値、すなわち通信回数を通信量として求める。

【0030】そして、次のステップS44で、通信回数配列におけるその移動候補のプログラムmと各実行器との要素の値の総和から、上記のステップS43で求めた移動先候補の実行器nの要素の値を減算し、減算結果に「1」を加えたものを、プログラムmを実行器nへ移動させたときの移動後通信量として見積もる。すなわちプログラムmと各実行器との間の通信量の総和からプログラムmと実行器nとの間の通信時間を減算することで、プログラムmを実行器nに移動させた後の、プログラムmと実行器n以外の他の実行器との間の通信時間を求める。また、プログラムmの移動前通信量として、通信回数配列におけるプログラムmと実行器i以外の他の実行器との要素の値の総和を求める。

【0031】なお、ステップS44の処理で、プログラムmを実行器nへ移動させた場合の移動後通信量を見積もる際に「1」を加算しているのは、プログラムmを実行器nへ移動させるときの通信量を加えるためである。

【0032】以上のようにしてプログラムmを実行器nへ移動させた場合の移動後通信量及び移動前通信量とを見積もったなら、同様にプログラムmを空いている他の実行器に移動させた場合の通信量を見積もる。

【0033】さらに、1つの移動候補のプログラムに関する移動後及び移動前通信量を見積もったなら、実行中

の他のプログラムについて同様に通信量の見積りを行う。そして、各プログラムの通信量を求めたなら、次にそれらのプログラムの通信時間を算出する。以下、通信時間見積り処理（通信時間算出手段に対応する）を図7（A）、（B）のフローチャートを参照して説明する。

【0034】まず、図7（A）のステップS51で、予め定められている、一定量の通信を行うときにかかる通信時間を通信時間係数として設定し、通信時間係数の初期化を行う。この実施例では、プログラムと実行器との間の通信時間が通信回数に比例するものとしているので、通信時間係数として「1」が設定される。

【0035】通信時間係数の初期化が終了したなら、図7（B）のステップS52で、前述した通信量見積り処理で求めた、移動候補のプログラムmを実行器nへ移動させた場合の通信量に通信時間係数を乗算して移動後通信時間を求める。本実施例では、通信量として通信回数をを用い、各実行器の通信時間係数を「1」としているの
10 ので、通信回数配列の各要素の通信回数がそのまま通信時間となる。なお、通信時間係数が実行器毎に異なる場合には、通信回数配列の各要素に、対応する実行器の通信時間係数を乗算して通信時間を求める。

【0036】そして、上記のようにして求めた移動後通信時間が、移動前通信時間より短くなるプログラムと実行器との組み合わせを抽出し、そのプログラムの実行を中断し、抽出した実行器をプログラムを再開すべき実行器として指定して、中断したプログラムのプログラムカウンタの値とデータとプログラム本体を制御器14に送信する。

【0037】制御器14は、実行器からプログラムを再開させるべき実行器を指定してプログラム情報が送られてきたなら、受信したプログラム情報を再開すべき実行器に送信し、プログラムを再開させる。

【0038】次に、本発明の負荷分散方法を、CPUの処理速度に比べてCPU間の通信時間の方が長い計算機システムに適用した場合について説明する。今、100個のCPUからなる計算機システムで、1つのCPU（実行器）が実行できるプログラム数の上限値が「900」、CPUにおける1つのプログラムの計算時間が「10」秒、1つのプログラムを他のCPUへ移動させる場合の移動時間が「300」秒で、全てのプログラムは1つのプログラムから生成されるものとする。

【0039】各CPUに均等にプログラムを割り当てる場合には、1000個のCPUにそれぞれ1つのプログラムを割り当てることとなるので、プログラムを生成したCPUから999個のCPUにプログラムを送信するための通信時間が「 300×999 秒」、計算時間が「10秒」であるので、計算全体にかかる時間は「 $300 \times 999 + 10 = 299,710$ 秒」となる。

【0040】これに対して、本発明の負荷分散方法を適用した場合、プログラムを他の実行器に移動させた場合

の計算時間より、1つのCPUにプログラムを集中させて場合の計算時間の方が短くなるので、1つのCPUに900個のプログラムを割り当て、残りの100個のプログラムを他のCPUに移動させることになる。その場合、100個のCPUにプログラムを割り当てるための通信時間が「 300×100 秒」、計算時間が「 10×900 秒」であるので、計算全体にかかる時間は「 $10 \times 900 + 300 \times 100 = 39,000$ 秒」となる。

【0041】従って、CPUの計算時間に比べてCPU間のデータの転送時間が長い場合には、各CPUに均等に負荷を割り当てる方法より、本発明の負荷分散方法の方が全体の計算時間を短縮することができる。

【0042】なお、上記実施例では、通信回数テーブル25に記憶されているプログラムの現在までの通信回数をもって通信量としているが、過去の一定期間の通信量をもって、そのプログラムの通信量としてもよい。また、通信回数ではなく、実際に送信されたデータ量をもって通信量としてもよい。

【0043】また、上記実施例では、通信時間係数を「1」として各実行器間の通信時間が等しい場合について説明したが、例えば、全てのプログラムの実行に先立って全ての実行器間で通信を行い、そのときの通信時間を各実行器の通信時間係数として用いてもよいし、移動すべきプログラムを抽出する際に、全ての実行器間、あるいはプログラムを移動させることのできる実行器との間で通信を行い、そのときの通信時間を実行器間の通信時間係数として用いてもよい。また、プログラムの行った過去全て、あるいは一定期間内の通信時間を通信量で割った値を通信を行った実行器間の通信時間係数として用いてもよい。あるいは、各実行器間の通信時間係数を予め個々に定めておいて、それらの値を通信時間係数として用いてもよいし、ネットワークにおける各計算機の接続状態に応じて定められた値を通信時間係数として用いてもよい。

【0044】

【発明の効果】本発明によれば、実行器での計算時間に比べて、実行器間の通信時間が長い場合に、プログラムを他の実行器に移動させたときの移動後の通信時間と、移動前の通信時間とを比較し、移動後の通信時間が短くなるようにプログラムを他の実行器に移動させることで、計算機システム全体の処理時間を短縮することができる。

【図面の簡単な説明】

【図1】本発明の原理ブロック図である。

【図2】実施例の計算機システムの構成図である。

【図3】実行器の構成図である。

【図4】実行器の動作を示すフローチャートである。

【図5】制御器の動作を示すフローチャートである。

【図6】（A）、（B）、（C）は通信量見積り処理のフローチャートである。

【図7】(A)、(B)、は通信時間見積り処理のフローチャートである。

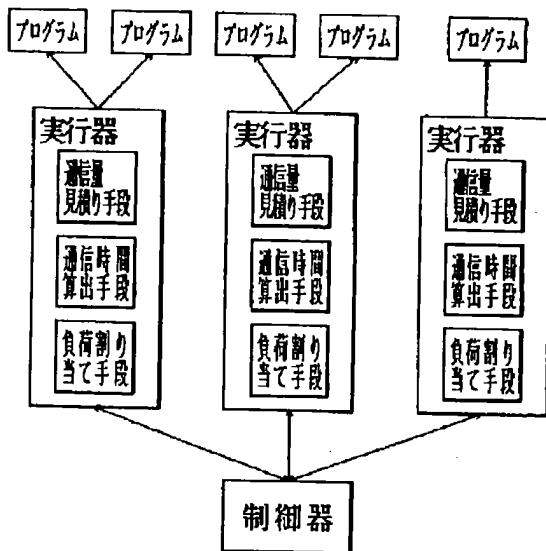
【符号の説明】

* 11~13 実行器
14 制御器

*

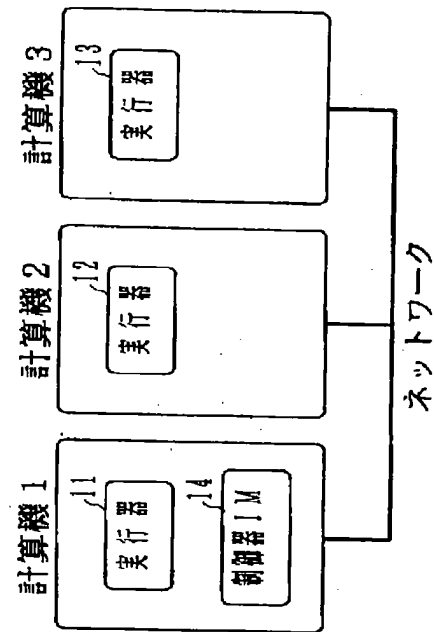
【図1】

本発明の原理ブロック図



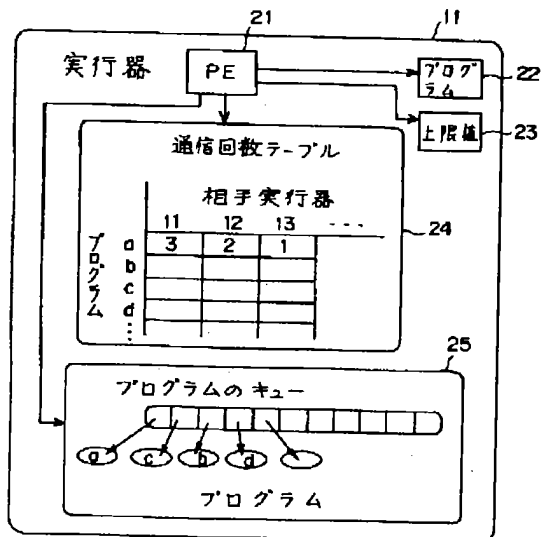
【図2】

実施例の計算機システムの構成図



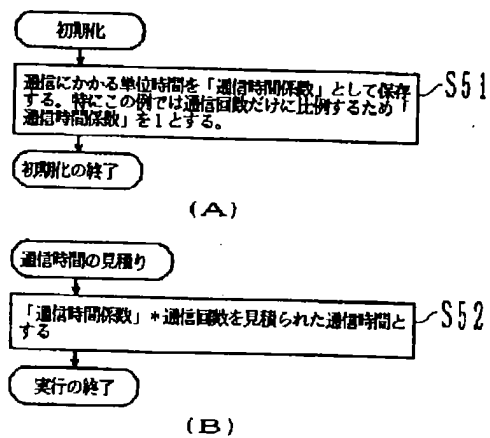
【図3】

実行器の構成図



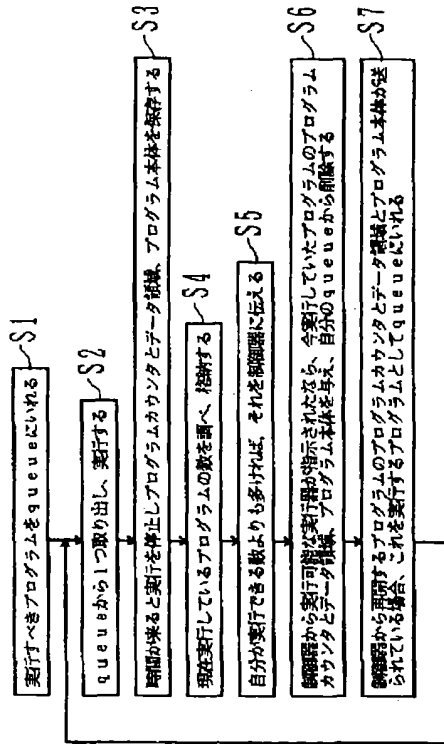
【図7】

通信時間見積り処理のフローチャート



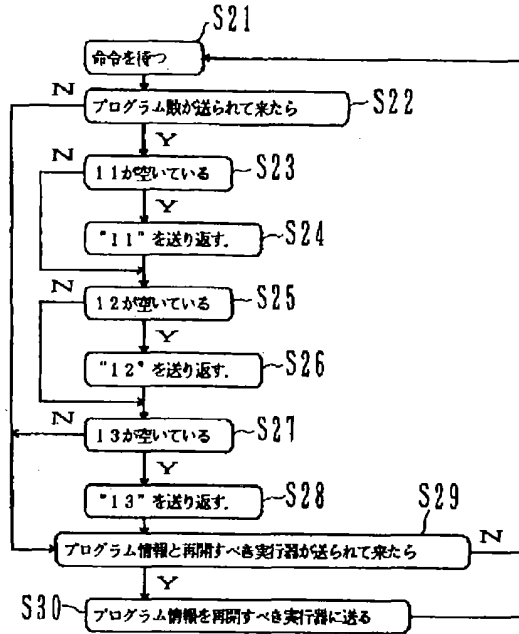
【図4】

実行器の動作を示すフローチャート



【図5】

制御器の動作を示すフローチャート



【図6】

通信量見積り処理のフローチャート

